Resolving outbreak dynamics using approximate Bayesian computation for stochastic birth-death models

METHOD ARTICLE

Resolving outbreak dynamics using approximate Bayesian computation for stochastic birth-death models [version 1; referees: 1 approved with reservations]

Jarno Lintusaari [1], Paul Blomstedt[1], Tuomas Sivula[1], Michael U. Gutmann[2], Samuel Kaski[1], Jukka Corander[3-5]

[1]Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University, Espoo, Finland
[2]School of Informatics, The University of Edinburgh, Edinburgh, UK
[3]Department of Biostatistics, University of Oslo, Oslo, Norway
[4]Infection Genomics, The Wellcome Trust Sanger Institute, Hinxton, UK
[5]Helsinki Institute for Information Technology HIIT, Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland

v1 **First published:** 25 Jan 2019, **4**:14 (https://doi.org/10.12688/wellcomeopenres.15048.1)
Latest published: 25 Jan 2019, **4**:14 (https://doi.org/10.12688/wellcomeopenres.15048.1)

Abstract

Earlier research has suggested that approximate Bayesian computation (ABC) makes it possible to fit simulator-based intractable birth-death models to investigate communicable disease outbreak dynamics with accuracy comparable to that of exact Bayesian methods. However, recent findings have indicated that key parameters such as the reproductive number *R* may remain poorly identifiable with these models. Here we show that the identifiability issue can be resolved by taking into account disease-specific characteristics of the transmission process in closer detail. Using tuberculosis (TB) in the San Francisco Bay area as a case-study, we consider a model that generates genotype data from a mixture of three stochastic processes, each with their distinct dynamics and clear epidemiological interpretation.

We show that our model allows for accurate posterior inferences about outbreak dynamics from aggregated annual case data with genotype information.

As a by-product of the inference, the model provides an estimate of the infectious population size at the time the data was collected. The acquired estimate is approximately two orders of magnitude smaller compared to the assumptions made in the earlier related studies, and much better aligned with epidemiological knowledge about active TB prevalence. Similarly, the reproductive number *R* related to the primary underlying transmission process is estimated to be nearly three-fold compared with the previous estimates, which has a substantial impact on the interpretation of the fitted outbreak model.

Keywords

Outbreak dynamics, Stochastic birth death process, Tuberculosis, Approximate Bayesian computation;

Open Peer Review

Referee Status: ?

	Invited Referees 1
version 1 published 25 Jan 2019	? report

1 **Jakub Voznica**, C3BI USR 3756 Institut Pasteur & CNRS, France
Anna Zhukova, C3BI USR 3756 Institut Pasteur & CNRS, France
Olivier Gascuel, C3BI USR 3756 Institut Pasteur & CNRS, France

Discuss this article

Comments (0)

Corresponding author: Jarno Lintusaari (jarno.lintusaari@aalto.fi)

Author roles: Lintusaari J: Formal Analysis, Investigation, Methodology, Writing – Original Draft Preparation; **Blomstedt P**: Formal Analysis, Writing – Original Draft Preparation, Writing – Review & Editing; **Sivula T**: Formal Analysis, Writing – Original Draft Preparation, Writing – Review & Editing; **Gutmann MU**: Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Kaski S**: Methodology, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Corander J**: Methodology, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by the Academy of Finland (Finnish Centre of Excellence in Computational Inference Research COIN) (grants 294238, 292334), the ERC (grant 742158), and the Wellcome Trust (grant 206194).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Lintusaari J, Blomstedt P, Sivula T *et al.* **Resolving outbreak dynamics using approximate Bayesian computation for stochastic birth-death models [version 1; referees: 1 approved with reservations]** Wellcome Open Research 2019, **4**:14 (https://doi.org/10.12688/wellcomeopenres.15048.1)

First published: 25 Jan 2019, **4**:14 (https://doi.org/10.12688/wellcomeopenres.15048.1)

Introduction

Birth-death processes are flexible models used for numerous purposes, in particular for characterizing spread of infections under the so called Susceptible-Infectious-Removed (SIR) formulation of an epidemic process[1]. Under circumstances where an outbreak of a disease occurs, but daily, weekly or even monthly incidence counts are not directly applicable or even available, the estimation of key epidemiological parameters, such as the reproductive number R, has to be based on alternative sources of information. This can be the case when the disease demonstrates large variability between the times of infection and onset, such as with *Mycobacterium tuberculosis*, or in retrospective analyses where all the information is no longer available. In such situations aggregate measures of the clusteredness of cases, for instance by genotype fingerprints, can be used as alternative source of information. The likelihood-based inference could provide an alternative to standard outbreak investigations relying solely on incident count data, but is often considerably more challenging.

As a solution to such a setting, Tanaka *et al.*[2] proposed fitting birth-death (BD) models to tuberculosis (TB) outbreak data using approximate Bayesian computation (ABC). Later on the same setting was used in numerous ABC studies while the ABC methodology was being developed [see e.g. 3–8]. Stadler[9] and Aandahl *et al.*[10] also tested the ABC procedure against an exact Bayesian inference method based on elaborate Markov Chain Monte Carlo (MCMC) sampling scheme. These investigations considered TB outbreak data from San Francisco Bay area originally collected by Small *et al.*[11], who reported results from extensive epidemiological linking of the cases, as well as the corresponding classical IS6110 fingerprinting genotypes. Such genetic data from the causative agent *Mycobacterium tuberculosis* are natural to characterize using the infinite alleles model (IAM), where each mutation is assumed to result in a novel allele in the bacterial strain colonizing the host. When lacking precise temporal information about the infection and the onset of the active disease, the numbers and sizes of genotype clusters can be used to infer the parameters of the BD model as shown by Tanaka *et al.*[2], Aandahl *et al.*[10].

Lintusaari *et al.*[12] demonstrated an issue with non-identifiability of R for the TB outbreak model in cases when both the birth and death rates were unknown in the underlying birth-death process. This was visible as a nearly flat approximate likelihood over the parameter space of R. Also it was found that in cases when R was identifiable, the acquired estimate was dependent on the assumed population size n. In the earlier investigations by Tanaka *et al.*[2] it was concluded that a large infectious population size $n = 10000$ was required for the BD simulator to produce similar levels of genetic diversity as observed in the San Francisco Bay data. Because it is unobserved, this assumption is difficult to justify while the acquired estimates will be dependent on it.

Here we introduce an alternative formulation of the BD model which resolves the identifiability issue of R[13].

The proposed model also does no longer require an assumption of the underlying infectious population size but provides an estimate for it as a by-product of the inference. The model incorporates epidemiological knowledge about the TB infection and disease activation processes by assuming that the observed genotype data represent a mixture of three birth-death processes, each with clearly distinct characteristics. The new formulation depends on a partly different parametrization for which estimates can be found from the literature. By evaluating the ABC inference results of our model in the light of the epidemiological information available from Small *et al.*[11], it is seen that both the significantly reduced infectious population size n and the increased R for the main driver component of the model make good sense. Our model thus provides a drastically changed interpretation of these parameters in comparison to the earlier studies.

In the new model we consider latent and active TB infections separately, as only the latter may lead to new transmission events. Transmission clusters are formed by a recent infection that rapidly progresses to an active TB and is spread further in the host population. Due to the rapid onset, the fingerprint of the pathogen remains the same in the transmission process and the patients consequently form an epidemiological cluster. If, on the other hand, the infection remains latent, the pathogen will undergo mutations and hence alters its fingerprint over the years[11]. By this and other epidemiologically motivated modelling choices we show that the model becomes identifiable. Due to the rather modest requirements for the available data and flexibility of modelling in ABC, our BD model can be applied to many similar settings beyond the case study considered in this article.

Model

The new model is based on the birth-death (BD) process where birth events correspond to an appearance of a new case with an active TB. A death event corresponds to any event that makes the existing host non-infectious, such as death, sufficient treatment, quarantine or relocation away from the community under investigation. The model incorporates two BD processes and one pure birth process that have a epidemiologically based interpretation. As in the standard BD process, the events are assumed to be independent of each other and to occur at specific rates. The time between two events is assumed to follow the exponential distribution specified by the rate of occurrence, causing the number of events to follow the Poisson distribution. The time scale considered here is one calendar year. The evolution of an infectious population is simulated by drawing events according to their rates.

Building upon the BD process, the simulated population carries auxiliary information. At birth, a case is assigned a cluster index that represents the specific genetic fingerprint of the pathogen and determines the cluster the case belongs to. The simulated output includes the cluster indexes that are recorded when cases become observed. Next we explain the model in more detail and notify differences to the model of Tanaka *et al.*[2].

First, we assume that observations are collected within a given time interval that matches the observed data. In the case of the San Francisco Bay data, the length of this interval is two years[11]. The observations are collected from the simulated process after a sufficient warm-up period, so that the process can be expected to have reached stable properties (exemplified in Figure 1).

A patient becomes observed in the study with probability p_{obs} when they cease to be infectious, i.e. when they undergo a death event in the simulation. Combining both being observed and ceasing to be infectious under the death event is based on the assumption that a typical patient is treated promptly after being diagnosed[14]. In contrast to the model of Tanaka *et al.*[2], there is then no separate observation sampling phase nor a prior estimate for the underlying population size.

Second, a burden rate parameter β is introduced to reflect the rate at which new active TB cases with a previously unseen fingerprint of the pathogen appear in the community. This is the pure birth process of the model and reflects the reactivations of TB from the underlying latently infected population and immigration. In the simulation, such cases receive a new cluster index that has not been assigned to any earlier case. Unlike Tanaka *et al.*[2], mutations are not explicitly modelled, but are assumed to occur during the latent phase of infection over the years[11].

Third, two distinct birth-death processes are introduced for cases that are either *compliant* or *non-compliant* with treatment. The birth-death processes are parametrized with birth rates τ_i, and death rates δ_i, where subscript $i = 1$ denotes the non-compliant population and $i = 2$ the compliant population. As noted in Small *et al.*[11], a significant factor behind the largest clusters in the observed data were non-compliant patients who stayed infectious for several months and belonged to subgroups under increased risk of rapid development of active TB due to conditions such as AIDS or substance abuse. Typical patients who are compliant with the therapy cease to be infectious relatively fast and do not transmit the disease as effectively before their diagnosis and treatment. Meta-analysis of typical time delays before diagnosis can be found from Sreeramareddy *et al.*[14].

We assume that a new TB case is non-compliant with therapy with probability p_1. At transmission (birth event) in the simulation this probability is used to determine the patient type of the new case. We also assume that the epidemic is in a steady state (Figure 1) by requiring that compliant cases have a reproductive number $R_2 = \tau_2/\delta_2 < 1$ below one and that the reproductive number R_1 of the non-compliant cases is constrained such that the population does not grow without limit. The steady state assumption is motivated by investigating the tuberculosis incidence counts in the United States during the data collection period[15]. We will next identify the subspace of the parameter values R_1 and R_2 that conform to this assumption.

Analysis of the model

Let subscript $i = 1$ denote a parameter of the non-compliant subpopulation and $i = 2$ the compliant subpopulation. The sizes of the subpopulations can be analyzed by investigating the parameters of the three birth-death processes in the proposed model. First we notice that the size of a subpopulation follows a compound birth-death process whose birth-rate is a linear function of the burden rate and the birth rates of the two subpopulations at their respective sizes. For instance the birth-rate of the non-compliant subpopulation is $p_1(\beta + \tau_1 n_1 + \tau_2 n_2)$ where n_1 and n_2 are the current sizes of the subpopulations and p_1 is the probability of a case being non-compliant. The corresponding death rate is $\delta_1 n_1$. Using this approach we can determine the balance sizes b_1 and b_2 of the subpopulations, meaning the values of n_1 and n_2 for which the birth and death rates of both of the subpopulations are equal. This corresponds to a state where the subpopulation sizes neither shrink or grow. The balance values b_2 and b_1 are obtained by solving the following set of linear equations:

$$\begin{aligned} \delta_1 b_1 &= p_1(\beta + \tau_1 b_1 + \tau_2 b_2), \\ \delta_2 b_2 &= p_2(\beta + \tau_1 b_1 + \tau_2 b_2), \end{aligned} \tag{1}$$

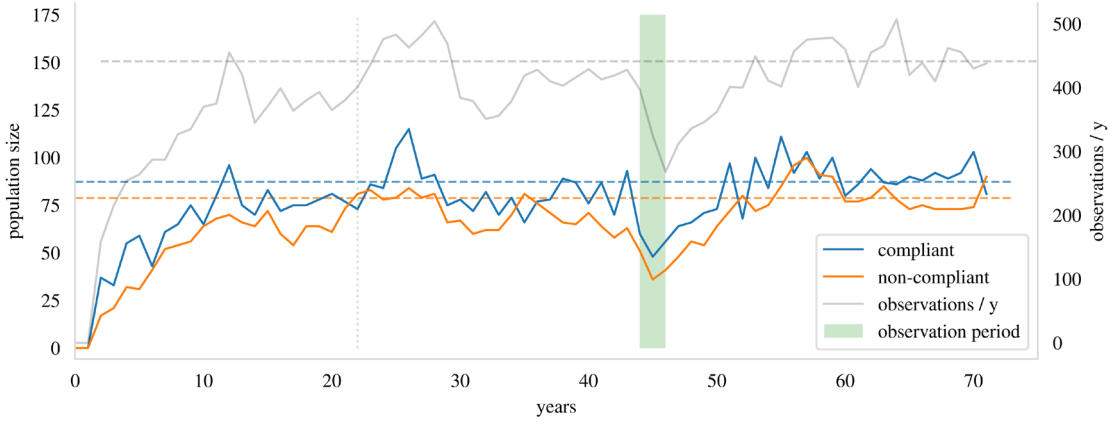

Figure 1. An illustration of simulated compliant and non-compliant populations as observed in the end of each year. The dashed lines are the balance values. The population sizes fluctuate around them after the process has matured. Both populations have surpassed their balance value at least once after 22 years. The observation period is the green patch. The grey line shows the number of observations that would have been collected within each year in the simulation. The number of observations from the observation period together with the clustering structure of the observations are used in the inference of the epidemiological parameters.

where $p_2 = (1 - p_1)$ is the probability of a new case being compliant. The linear equations yield the following solution

$$b_1 = \frac{p_1 \beta \delta_2}{\delta_2 \delta_1 - p_2 \tau_2 \delta_1 - p_1 \tau_1 \delta_2} \qquad (2)$$
$$b_2 = \frac{b_1(\delta_1 - p_1 \tau_1) - p_1 \beta}{p_1 \mathrm{t}_2}.$$

Given this solution, the balance values b_1 and b_2 exist when

$$R_1 < 1/p_1, \qquad (3)$$
$$R_2 < (1 - p_1 R_1)/p_2.$$

Assuming for instance that $p_2 = .95$[11], this translates to $R_1 < 20$.

Equation 2 allows also one to approximate the mean number of observed cases per year by defining the approximation as

$$\hat{n}_{obs} = p_{obs}(\delta_2 b_2 + \delta_1 b_1). \qquad (4)$$

Figure 1 illustrates how the population sizes fluctuate near their balance values in the simulation after a sufficient warm-up period.

Parameter inference

Approximate Bayesian computation was used to carry out the parameter inference due to the unavailability of the likelihood function. This is the same approach as used by Tanaka *et al.*[2] with the original model. The result will be a sample from the approximate posterior distribution $\tilde{p}(R_1, t_1, R_2, \beta \mid y_0)$ [see e.g. 16].

We used the Engine for Likelihood-Free Inference (ELFI) software[17] to perform the inference. We sampled 1000 parameter values with rejection sampling from a total of 6M simulations. A visualization of the ELFI model can be found from Figure S1 in the Extended Data. The observed data are available in the article of Small *et al.*[11]. Furthermore we have released the source code of the simulator and the corresponding ELFI model in GitHub that allow a replication of this study[13].

Priors

We set priors over the burden rate β, reproductive numbers R_1 and R_2, and the net transmission rate $t_1 = \tau_1 - \delta_1$ of the non-compliant population. For the compliant population the death rate is fixed to an estimate $\delta_2 = 5.95$[14], the total delay estimate] that can be transformed to a net transmission rate via $t_2 = \delta_2(R_2 - 1)$. Based on the details in Small *et al.*[11] describing the San Francisco Bay area TB data, there were 585 confirmed cases of TB of which 487 were included in the study. To account for the cases that were not included in the study, we fix the probability of becoming observed to $p_{obs} = 0.8$. The probability of a new case being non-compliant was set to the estimate $p_1 = 0.05$ [11, page 1708].

The burden rate β is given an informative prior that is able to produce a sufficient number of clusters with respect to the observed data. Specifically, we set

$$\beta \sim \mathrm{N}(200, 30). \qquad (5)$$

The net transmission rate t_1 is given a uniform prior over a large interval from 0 to 30. Given the solution in Equation 3, the reproductive numbers R_1 and R_2 are given uniform prior over subspace that ensures the process has a steady state. More specifically

$$R_1 \sim \mathrm{Unif}(1.01, 20), \qquad (6)$$
$$R_2 \mid R_1 \sim \mathrm{Unif}(0.01, (1 - 0.05 \cdot R_1)/0.95),$$
$$t_1 \sim \mathrm{Unif}(0.01, 30),$$

To optimize the computation given the observed data, we set the following additional constraints:

$$\hat{n}_{obs} < 350, \qquad (7)$$
$$\tau_1 < 40.$$

These constraints were checked to have a negligible effect on the acquired estimates. They however prevented simulations with extremely unlikely parameter values thus saving considerable amount of computation time. Effectively due to the constraints, the values of R_1 were smaller than 15. Figure 2 shows samples drawn from the priors under these constraints.

Summary statistics

The summary statistics used in earlier approaches [see e.g. 2, 12] are not directly applicable to the proposed model. This is due to the differences between the models that cause for example the number of observations in the sample to vary rather than being fixed. However, the earlier summaries still provide a good starting point for developing a more comprehensive set of summaries.

We use the following eight summary statistics which aim to capture meaningful properties of the observed data given the new model. The first summary is the *number of observations* which is allowed to vary in the new model. Five of the summaries are related to the clustering structure: *the total number of clusters, the relative number of singleton clusters, the relative number of clusters of size two, the size of the largest cluster and the mean of the successive difference in size among the four largest clusters* (Table 1). These were chosen in an attempt to emphasize the more stable properties of the clustering structure. For instance there is a substantial number of singleton and size two clusters compared to other cluster sizes. The relative number is used to remove the effect of variability in the numbers of observations and clusters between simulations.

The remaining two summaries are related to the observation times of the largest cluster. Observation times were not included in the earlier approaches and proved to be useful in identifying the net transmission rate t_1. These are *the number of months from the first observation to the last* and *the number of months when at least one observation was made*. This data could be extracted from Figure 2 in Small *et al.*[11]. With these summaries we aim to capture the span and rate at which transmissions occur.

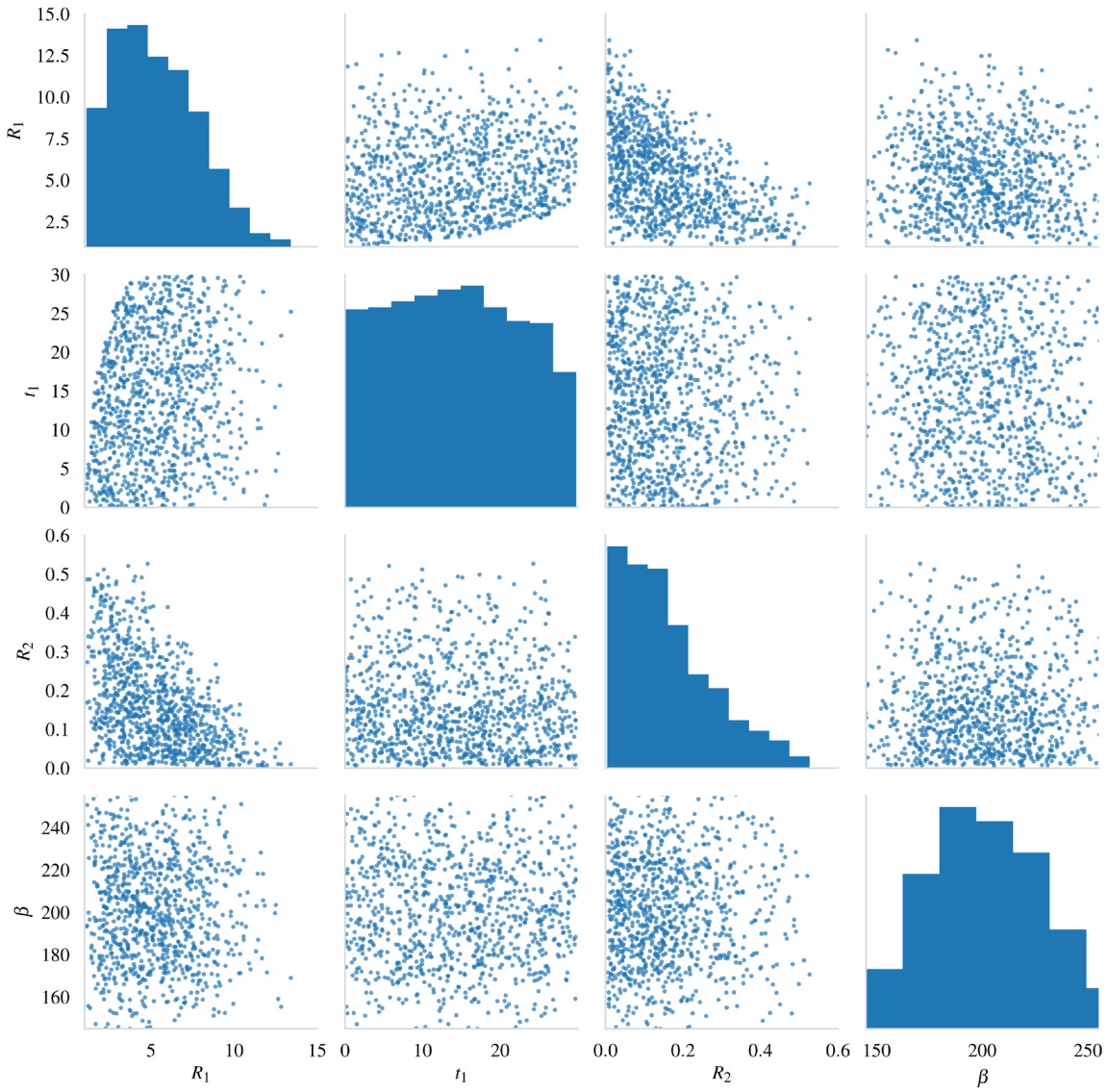

Figure 2. A scatter matrix of samples from the prior.

Table 1. The summary statistics, their weights, and the values of the summary statistics for the observed data y_0.

Summary statistic	Explanation	Weight	y_0
n_{obs}	Number of observations.	1	473
$n_{clusters}$	Number of clusters.	1	326
r_{c1}	Relative number of singleton clusters. Computed as $r_{c1} = n_{c1}/n_{obs}$ where n_{c1} is the number of clusters of size 1. The value of r_{c2} is computed likewise.	100/0.60	0.60
r_{c2}	Relative number of clusters of size 2.	100/0.04	0.04
largest	Size of the largest cluster.	2	30
mean_largest_diff	Mean of the successive difference in size among the four largest clusters.	10	6.67
month_period	Number of months from the first observation to the last in the largest cluster.	10	24
obs_months	The number of months that at least one observation was made from the largest cluster.	10	17

It is good to note that the summaries chosen here do not consider global sufficiency [see e.g. 18]. In cases where the dataset is very different from the San Francisco data, a modified set of summaries should probably be considered. The distance function is the Euclidean distance between the weighted summary statistics of the observed and simulated data (Table 1). The weights were chosen to adjust and even up differences in the magnitudes of the different summaries. The chosen summary statistics and weights were found to perform well in the evaluation of the model in Subsection . The resulting threshold for the acquired sample was $\epsilon = 31.7$ with the smallest distance being 12.5.

Results

Figure 3 shows a sample of 1000 values from the joint approximate posterior distribution $\tilde{p}(R_1, t_1, R_2, \beta \mid y_0)$. The pairwise sample clouds seem reasonably concentrated and are away from the edges of the axes and inside the support of the prior (Figure 2). The histograms and scatter plots look rather normally shaped, the only minor exception being the net transmission rate of the non-compliant population t_1, that has a slight tail towards high values. A visual comparison of the posterior against the prior together with the above observations suggest that the model is identifiable for the San Francisco dataset.

The posterior means, medians and 95% credible intervals are given in Table 2. The means and medians are close to each other indicating symmetry of the posterior distributions. The t_1 has the largest discrepancy due to its small tail mentioned above.

Evaluating the model identifiability

To further evaluate the reliability of the acquired estimates, we compute the mean and median absolute errors (MAE and MdAE) of the mean, and the coverage property (Wegmann *et al.*, 2009),

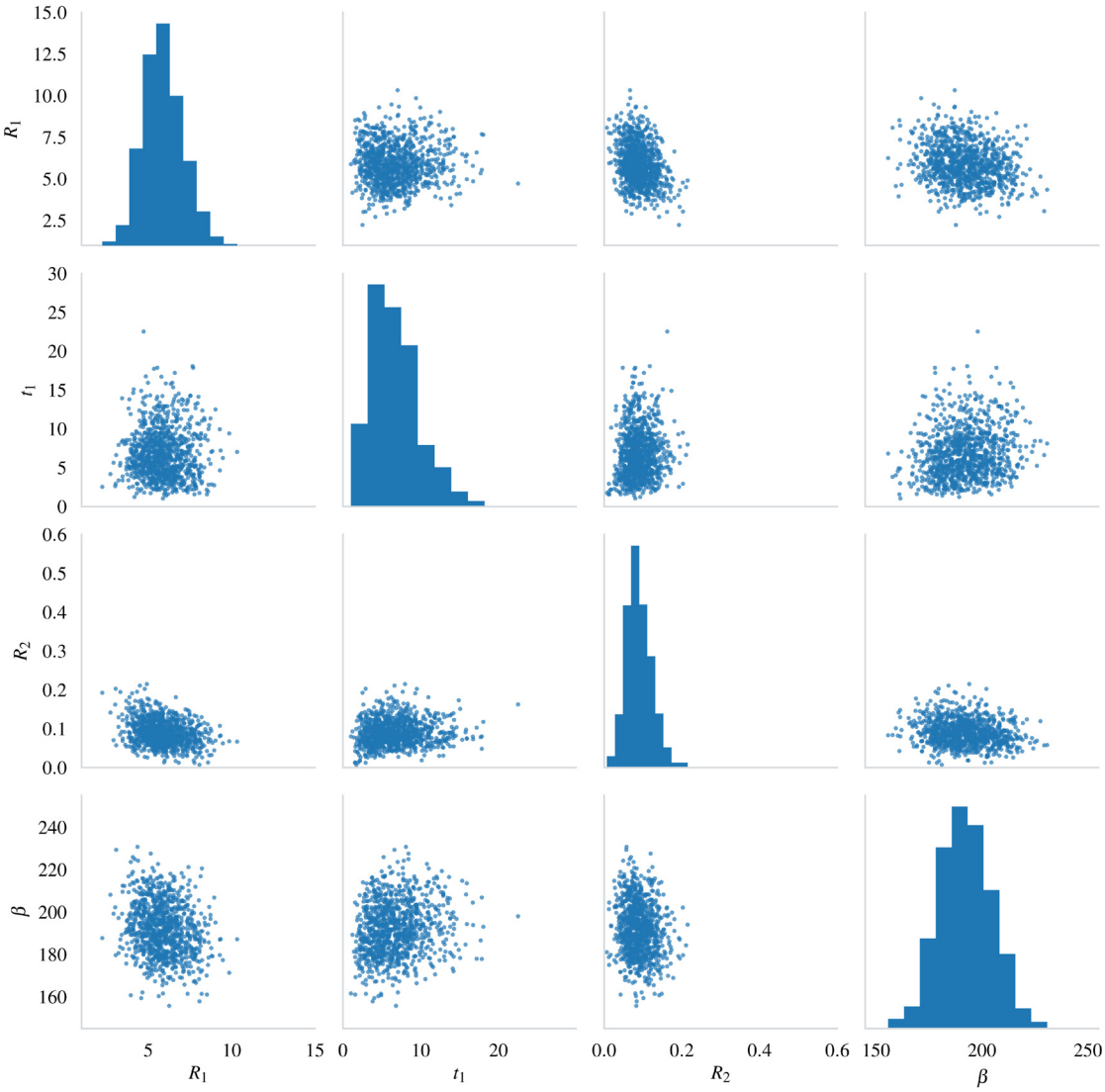

Figure 3. Posterior sample of size 1000 from the approximate posterior distribution $\tilde{p}$ (R_1, t_1, R_2, β | y_0) plotted as a scatter matrix. Compare to the prior in Figure 2.

with 1000 synthetic observations from the posterior with known parameter values. These results include the ABC approximation error [see e.g 16] caused by the summary statistics and the threshold of 31.7.

Table 3 lists the MAE and MdAE with the 95% error upper percentile. These are useful in quantifying how much the estimate deviates from the actual value on average. The burden rate β and the reproductive number of the non-compliant population R_1 have the smallest relative MAEs, 4.0% and 14.9%, respectively. The reproductive number R_2 of the compliant population and the net transmission rate t_1 of the non-compliant population have MAEs of 29.5% and 44.2%. The MAE of the latter seems rather high. Also the 95% percentile (Table 3) indicates that in 5% of the trials the error was substantial. Investigating the issue further showed that for some of the synthetic datasets, the net transmission rate parameter t_1 was not identifiable, meaning that the synthetic data in those cases was not informative enough to produce a clear mode for the parameter. Also R_2 suffered slightly from the same issue. This kind of situation where some of the synthetic datasets turn out uninformative is a rather common occurrence in cases where there is little data available. Because of these exceptions, MdAE might be a more appropriate measure as it is not as much influenced by the results of the non identifiable datasets in the trials. Relative MdAE errors were 21.9% and 32.1% respectively. Figure 4 visualizes the estimated values against their actual values for each of the parameters.

The coverage property[19] is used to assess the reliability of the inference by checking whether the spreads of the acquired posterior distributions are accurate. Given a critical level α, the true parameter value should be outside the $(1-\alpha)$ credible interval of the posterior with probability α. The estimated α-values from 1000 marginal posteriors with known true parameter values were satisfactory (Figure 5). For the critical level $\alpha = .05$ the estimated α-values were $(\alpha_{R_2}, \alpha_{R_1}, \alpha_{\beta}, \alpha_{t_1}) = (.03, .03, .02, .04)$. The overall performance with different α was similar to this case in the sense that α_{β} suffered from a larger error compared to the estimates for the other parameters (Figure 5).

Table 2. Posterior summaries.

Parameter	Mean	Median	95% CI
R_1	5.88	5.79	(3.68, 8.16)
t_1	6.74	6.25	(1.57, 12.9)
R_2	0.09	0.09	(0.03, 0.15)
β	192	192	(170, 216)

Table 3. Mean and Median Absolute Errors in 1000 trials with synthetic data from the posterior.

Parameter	MAE	Relative MAE[1]	MdAE	Relative MdAE	95% percentile
R_1	0.85	14.9%	0.72	12.6%	2.00
t_1	2.68	44.2%	1.98	32.1%	7.66
R_2	0.024	29.5%	0.018	21.9%	0.07
β	7.6	4.0 %	6.1	3.1%	19.8

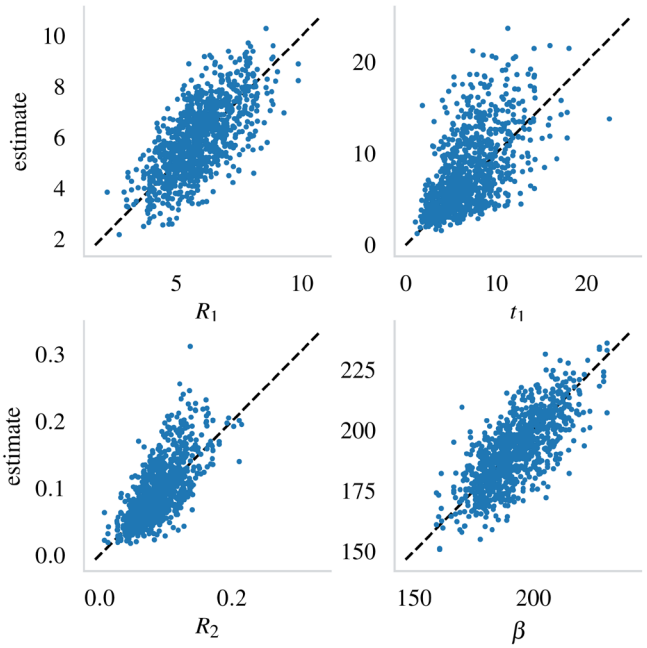

Figure 4. The estimates from the 1000 trials plotted against their true values. The black dashed line shows the 1:1 correspondence.

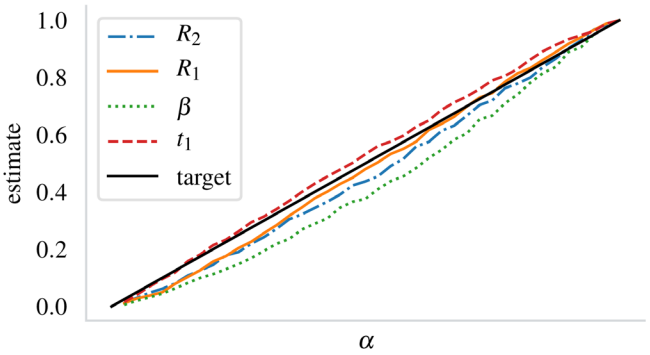

Figure 5. Mean estimates for the critical level α at different levels. The estimates are computed from 1000 synthetic datasets from the posterior. For the reference, the estimates for α = .05 were (.030, .028, .020, .041) in the same order as in the legend.

Discussion

We have proposed a stochastic birth-death model extending from several previous articles examining the use of simulator-based inference for the spread of active TB within a community. Outbreaks of TB are characterized by epidemiologically linked clusters of patients with active TB that emerge within a relative

short time interval. The construction of the extended model was motivated by several epidemiological observations made by Small *et al.*[11] concerning the San Francisco Bay transmission cluster data. Each of the largest clusters were largely formed by a non-compliant patient. In the largest cluster such a patient apparently infected 29 additional patients. The earlier approach[2,10] suffered from inability to reproduce these large clusters with an appropriate level of heterogeneity in the cluster sizes without a prior assumption of a very large underlying infectious population size (in the order of 10000)[2,12]. Based on epidemiological knowledge about TB such a large infectious population size is unlikely to have existed in the study region during the observation period. Furthermore it was shown that this assumption has a considerable effect on the estimate of the reproductive number R.

Under our new model, a prior estimate of the infectious population size is not needed. Instead the new model has a different parametrization for which estimates can be found from the literature. As a by-product of the inference, the model also yields estimates for the infectious population size at the end of the data collection period. For the San Francisco Bay data the mean and median sizes of the compliant subpopulation were 48.4 and 48 respectively, and 13.5 and 11 for the non-compliant subpopulation.

The reproductive numbers R_1 and R_2 of the subpopulations represent the average number of infections that rapidly progress to active TB, caused by a single already infectious case. This counting therefore excludes infections remaining latent, which are instead indirectly captured via the burden rate parameter β. We estimate that the reproductive number of the non-compliant patients is $R_1 = 5.88$ with the 95% credible interval (CI) (3.68, 8.16) (Table 2). The estimate is nearly three-fold compared to the estimate of 2.10 in Aandahl *et al.*[10] with the same data, which provided a single estimate for the whole infectious population without considering differences between patient types. The larger value seems reasonable in explaining the formation of large clusters within a short time. The reproductive number of compliant cases is estimated to be $R_2 = 0.09$ with a 95% CI (0.03, 0.15).

The ability of the proposed model to estimate R_1 and R_2 together with the population size follows from several important changes in the proposed model. One of them is that observations are collected during the observation period that matches the length of the actual observation period. In the original model observations were collected as a snapshot at a single point of time which required that all patients in a large cluster had to be infectious at the same time. However, the observed counts are in reality a result of observations made over time as the local outbreak evolves and with some patients having separate infectious periods. Figure 2 in Small *et al.*[11] shows how the patients were diagnosed at different times over the observation period. Another factor is the inclusion of non-compliant patient type in the model which more closely represents the description of Small *et al.*[11] and naturally enables the formation of heterogeneity in the cluster sizes.

It should be noted that being compliant or non-compliant are thought to characterize the type of a patient and the model decides this at the time of the birth event. In reality, non-compliant patients are often diagnosed (i.e. observed) earlier compared to when they cease to be infectious, which implies that the simulator model deviates slightly from typical observation processes in this respect. However, considering that this discrepancy applies to only roughly 5% of all the observed cases, we do not expect any sizeable bias to arise from this assumption. Furthermore, the summary statistics used do not consider exact diagnosis times but rather just the span and the rate at which they occur.

The model identifiability was found to be satisfactory for the San Francisco Bay dataset (Figure 3). The average error in the estimate of R_1 with the proposed method is evaluated to be 14.9% (0.85 in absolute terms, Table 3). The same for R_2 is 29.5% (0.024 absolute), although the median error (21.9%, 0.018 absolute) is probably a more reasonable value due to the reasons discussed earlier. The coverage property analysis[19] suggests that the credible intervals provided by the model are sensible. For future work it would be also interesting to evaluate the sensitivity of the model to other possible choices for the literature based parameter estimates.

As the IS6110 typing remains in epidemiological use, despite advances in whole-genome sequencing of TB isolates, our model could be used for investigations in particular in middle and low income countries, where the TB burden is often also highest. For example, the estimates for the epidemiological parameters could be used to gain insight to the relative efficacy of the control programs across multiple communities. Given the apparent success by which the non-identifiability issue for R and the assumption of *a priori* known infectious population size were resolved by extending the BD model by relevant and often available epidemiological knowledge, it would be interesting to generalize the approach in the future to other pathogens for which the sampling process or other factors render the simulator-based inference as the most promising estimation method.

Data availability

The observed data are available in the article of Small *et al.*[11]. Figure S1 is available from https://doi.org/10.6084/m9.figshare.7578728.v1.

Software availability

Source code available from: https://github.com/jlintusaari/tb-model

Archived source code at time of publication: https://doi.org/10.5281/zenodo.2540933

License: BSD 3-Clause.

Author contributions

JL was the principal writer of the manuscript, designed and implemented the proposed model and carried out experiments.

PB and TS participated in the writing of the paper (results section) and carried out experiments. MG participated in the writing of the paper (methods section) and in the design of the proposed model. SK and JC wrote partly major parts of the paper (introduction, discussion, methods), directed the study and participated in the design of the proposed model.

Grant information

This work was supported by the Academy of Finland (Finnish Centre of Excellence in Computational Inference Research COIN) (grants 294238, 292334), the ERC (grant 742158), and the Wellcome Trust (grant 206194).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

We acknowledge the computational resources provided by the Aalto Science-IT project.

An earlier version of this article can be found on bioRxiv (DOI: https://doi.org/10.1101/215533)

References

1. Anderson RM, May RM: **Infectious Diseases of Humans: Dynamics and Control**. Oxford University Press, 1992.
Reference Source
2. Tanaka MM, Francis AR, Luciani F, *et al.*: **Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data.** *Genetics.* 2006; **173**(3): 1511–1520.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Sisson SA, Fan Y, Tanaka MM: **Sequential Monte Carlo without likelihoods.** *Proc Natl Acad Sci U S A.* 2007; **104**(6): 1760–1765.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Blum MGB: **Approximate Bayesian computation: A nonparametric perspective.** *J Am Stat Assoc.* 2010; **105**(491): 1178–1187.
Publisher Full Text
5. Fearnhead P, Prangle D: **Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate bayesian computation.** *J R Stat Soc Series B Stat Methodol.* 2012; **74**(3): 419–474.
Publisher Full Text
6. Del Moral P, Doucet A, Jasra A: **An adaptive sequential Monte Carlo method for approximate Bayesian computation.** *Stat Comput.* 2012; **22**(5): 1009–1020.
Publisher Full Text
7. Baragatti M, Grimaud A, Pommeret D: **Likelihood-free parallel tempering.** *Stat Comput.* 2013; **23**(4): 535–549.
Publisher Full Text
8. Albert C, Künsch HR, Scheidegger A: **A simulated annealing approach to approximate Bayes computations.** *Stat Comput.* 2015; **25**(6): 1217–1232.
Publisher Full Text
9. Stadler T: **Inferring epidemiological parameters on the basis of allele frequencies.** *Genetics.* 2011; **188**(3): 663–672.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Aandahl RZ, Stadler T, Sisson SA, *et al.*: **Exact *vs.* approximate computation: reconciling different estimates of *Mycobacterium tuberculosis* epidemiological parameters.** *Genetics.* 2014; **196**(4): 1227–1230.
PubMed Abstract | Publisher Full Text | Free Full Text
11. Small PM, Hopewell PC, Singh SP, *et al.*: **The epidemiology of tuberculosis in San Francisco. A population-based study using conventional and molecular methods.** *N Engl J Med.* 1994; **330**(24): 1703–1709.
PubMed Abstract | Publisher Full Text
12. Lintusaari J, Gutmann MU, Kaski S, *et al.*: **On the Identifiability of Transmission Dynamic Models for Infectious Diseases.** *Genetics.* 2016; **202**(3): 911–918.
PubMed Abstract | Publisher Full Text | Free Full Text
13. Lintusaari J: **jlintusaari/tb-model: Publication (Version v1.0).** *Zenodo.* 2019.
http://www.doi.org/10.5281/zenodo.2540933
14. Sreeramareddy CT, Panduru KV, Menten J, *et al.*: **Time delays in diagnosis of pulmonary tuberculosis: a systematic review of literature.** *BMC Infect Dis.* 2009; **9**: 91.
PubMed Abstract | Publisher Full Text | Free Full Text
15. CDC: **Reported Tuberculosis in the United States 2016**. 2017.
Reference Source
16. Lintusaari J, Gutmann MU, Dutta R, *et al.*: **Fundamentals and Recent Developments in Approximate Bayesian Computation.** *Syst Biol.* 2017; **66**(1): e66–e82.
PubMed Abstract | Publisher Full Text | Free Full Text
17. Lintusaari J, Vuollekoski H, Kangas-rääsiö A, *et al.*: **Elfi: Engine for likelihood-free inference.** *J Mach Learn Res.* 2018; **19**(16): 1–7.
Reference Source
18. Nunes MA, Balding DJ: **On optimal selection of summary statistics for approximate Bayesian computation.** *Stat Appl Genet Mol Biol.* 2010; **9**(1): Article34.
PubMed Abstract | Publisher Full Text
19. Wegmann D, Leuenberger C, Excoffier L: **Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood.** *Genetics.* 2009; **182**(4): 1207–18.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Referee Status: ?

Version 1

Referee Report 13 February 2019

https://doi.org/10.21956/wellcomeopenres.16417.r34679

? **Jakub Voznica** , **Anna Zhukova** , **Olivier Gascuel**
Unité Bioinformatique Evolutive, C3BI USR 3756 Institut Pasteur & CNRS, Paris, France

Article summary

The article describes a new model of TB outbreak in San Francisco Bay area that overcomes the non-identifiability/dependency on the assumed population size of the reproductive number R in the generic birth-death-mutation model by Tanaka *et al.* The new model considers two compartments, for compliant and non-compliant subpopulations, and combines two birth-death processes (for each of the compartments) with a pure-birth process that creates new TB transmission clusters (i.e. a new individual with a new RFLP pattern that is further transmitted). This pure-birth process replaces mutation in Tanaka's model and corresponds to migration or reactivating of a latent TB. The rate corresponding to the pure-birth process is referred as the burden rate. At each (non-burden) birth event (i.e. TB transmission) the compartment of the newly infected individual is assigned to non-compliant or compliant with the probability p1 or (1 - p1) correspondingly. At each death (i.e. becoming non-infectious) event the individual is sampled with the probability p-obs.

Overall, the proposed model has 7 parameters: the burden rate, 2 birth rates, 2 death rates, and 2 probabilities (p1 and p-obs). However, 3 of them (compliant death rate, p-obs and p1) were fixed based on the estimates from the literature, therefore leaving 4 parameters to be estimated, expressed in terms of two reproductive numbers, i.e. birth to death rate ratios for the corresponding compartments, the non-compliant net transmission rate (difference between the birth and the death rates), and the burden rate. Priors and additional constraints on the rates were set to avoid biological meaningless of the simulations.

The simulator was implemented for the proposed model and parameter estimation was performed for the data collected in SF Bay area in 1991-92 (Small et al.) with ABC, based on 1000 parameter values sampled with rejection from 6M simulations, using 8 (weighted) summary statistics:

1. the number of observations
2. the total number of clusters
3. the relative number of singleton clusters
4. the relative number of clusters of size two
5. the size of the largest cluster
6. the mean of the successive difference in size among the four largest clusters
7. the number of months from the first observation to the last
8. the number of months when at least one observation was made.

The new model not only allowed for estimation of the aforementioned parameters (posteriors are well concentrated within but far from the edges of the priors) but also of the balance subpopulation sizes (at the equilibrium state when infected subpopulations neither shrink nor grow). The estimates differ from those done with the birth-death-mutation model, and are potentially better aligned with the epidemiological knowledge on TB in the area.

The coverage property (accuracy of the spread of the acquired posterior) of the estimator was further tested on 1000 parameter values drawn from the posterior, giving satisfactory results for the critical level of .05 (the true parameter values were outside of the .95 credible interval of the posterior with probability less than .05).

General comments

The article reads well, the model, rationale behind it, its assumptions and advantages over the previous TB model are explained in a clear and convincing way. It is a valuable addition to TB research, and we believe that the article should be accepted.

Having little knowledge on TB (but on ABC), we feel like the article could benefit from a more detailed discussion of the obtained estimates. For example, is there any literature/other data supporting the estimated subpopulation sizes?

We also point out a few technicalities that could be explained in more detail (see below).

Technical comments

A flow diagram of the model could facilitate the model understanding for the reader.

Additional sensitivity analysis of the model while varying pre-fixed parameter values (of compliant death rate, p-obs and p1) might add confidence in author's findings.

Page 4: *"The observations are collected from the simulated process after a sufficient warm-up period, so that the process can be expected to have reached stable properties (exemplified in Figure 1)."*
In Figure 1 the warm-up seems to be achieved already after 15 years, however the observation period is chosen around 45 years, where there is a drop of population sizes. Is it a coincidence? How is the start of the observation period selected?

Page 5: *"We used the Engine for Likelihood-Free Inference (ELFI)..."*
The authors might detail what kind of inference was used: Is it a pure distance/rejection-based approach? Or do you use some regression tool, random forest, LASSO, neural network or other? How was the technique selected?

Page 5: *"Based on the details in Small et al. describing the San Francisco Bay area TB data, there were 585 confirmed cases of TB of which 487 were included in the study. To account for the cases that were not included in the study, we fix the probability of becoming observed to p-obs = 0.8"*
If we understand correctly the p-obs is calculated as 487/585, but what about potentially unknown cases of TB in the SF Bay area? Is it assumed that all the existing TB cases are known?

Page 5: It is not very clear why these particular summary statistics were selected, e.g. *"the mean of the successive difference in size among the four largest clusters"*

Why not 3 or 5, etc.? Were for example other statistics tested, which performed worse?
The name of the last statistic (*"the number of months when at least one observation was made"*) is rather confusing. In table 1 it has a slightly different name: *"the number of months that at least one observation was made from the largest cluster"*. Does it mean *the time when the first observation from the largest cluster was made*?

Page 7: *"The chosen summary statistics and weights were found to perform well in the evaluation of the model in Subsection ."*
The subsection number is missing.

Page 7: *"The resulting threshold for the acquired sample was $\epsilon = 31.7$ with the smallest distance being 12.5."*
How were the threshold and distance values selected?

Is the rationale for developing the new method (or application) clearly explained?
Yes

Is the description of the method technically sound?
Yes

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.